JVI Accepts, published online ahead of print on 15 May 2013
J. Virol. doi:10.1128/JVI.01225-12
Copyright © 2013, American Society for Microbiology. All Rights Reserved.

JVI Accepts published online ahead of print

HIV Populations are Large and Accumulate High Genetic Diversity in Nonlinear Fashion

Frank Maldarelli[1], Mary Kearney[1], Sarah Palmer[1*], Robert Stephens[2], JoAnn Mican[3], Michael A. Polis[3], Richard T. Davey[3], Joseph Kovacs[4], Wei Shao[2], Diane Rock-Kress[3], Julia A Metcalf[3], Catherine Rehm[3], Sarah E. Greer[5], Daniel L. Lucey[6], Kristen Danley[1], Harvey Alter[5], John W. Mellors[7], John M. Coffin[1,8]

[1]HIV Drug Resistance Program, NCI-Frederick, NIH, Frederick MD; [2]ISP/Advanced Biomedical Computing Center, SAIC, Frederick MD, Laboratory of Immunoregulation, NIAID, NIH,Bethesda MD, [4]Department of Critical Care, NIH, Bethesda MD, [5]Department of Transfusion Medicine, NIH, Bethesda MD, [6]Division of Infectious Diseases, Washington Hospital Center, Washington D.C., [7]Division of Infectious Diseases, University of Pittsburgh, Pittsburgh PA, [8]Department of Molecular Biology and Microbiology, Tufts University, Boston MA.

*Present Address: Westmead Millennium Institute for Medical Research, Sydney AU

Running Title: HIV populations are large, diverse and slow to change
Keywords: HIV, replicating population size, resistance, population genetics

JVI Accepts published online ahead of print

Abstract

HIV infection is characterized by rapid and error-prone viral replication resulting in genetically diverse virus populations. The rate of accumulation of diversity and the mechanisms involved are under intense study to provide useful information to understand immune evasion and the development of drug resistance. To characterize the development of viral diversity after infection, we carried out an in-depth analysis of single genome sequences of HIV *pro-pol* to assess diversity and divergence, and to estimate replicating population sizes in a group of treatment naive HIV-infected individuals sampled at single (N=22) or multiple, longitudinal time points (N=11). Analysis of single genome sequences (SGS) revealed non-linear accumulation of sequence diversity during the course of infection. Diversity accumulated in recently infected individuals at rates 30-fold higher than in patients with chronic infection. Accumulation of synonymous changes accounted for most of the diversity during chronic infection. Accumulation of diversity resulted in population shifts, but the rates of change were slow relative to estimated replication cycle times, consistent with relatively large population sizes. Analysis of changes in allele frequencies revealed effective population sizes that are substantially higher than previous estimates of approximately 1000 infectious particles/infected individual. Taken together, these observations indicate that HIV populations are large, diverse, and slow to change in chronic infection and that the emergence of new mutations, including drug resistance mutations, is governed by both selection forces and drift.

JVI Accepts published online ahead of print

Introduction

Infection with human immunodeficiency virus type 1 (HIV) results in lifelong persistent infection. In most cases, HIV infection results from expansion of a single or limited number of viral variants (24, 41, 49, 59), producing an initially uniform virus population. From the time of infection, HIV genetic diversity emerges as a function of mutation, drift, recombination, selection and population size. Early in infection, genetic diversity increases in a linear fashion (22), at rates somewhat lower than that predicted by the rapid (generation time 1-2 d) and error prone replication program (with the unselected reverse transcription mutation rate (3-5 x 10^{-5} mutations/base/replication cycle,31). HIV variants emerge with mutations at a number of positions; distribution of genetic distances in these early populations largely approximates a Possion distribution, suggesting that, in general, sites undergo mutation at random. Emergence of variants with mutations at specific CTL sites is relatively frequent, and suggests that although mutations may occur at random, individual variants emerge as escape mutations (22, 24). These data (22, 24) demonstrate a strong role for both mutation and selection in the formation of initial populations in infected individuals. After years of infection, substantial genetic diversity accumulates, and this highly diverse population can rapidly respond to selective pressures, facilitating immune escape, and resistance to antiviral drugs. Understanding how new mutations emerge and become fixed in HIV populations is critical to designing effective strategies for the prevention and suppression of these sequelae (2, 8, 9, 15, 16, 33, 37, 42).

Although much has been learned regarding establishing HIV infection in vivo, critical gaps in our understanding persist that limit our understanding of the dynamics of HIV populations and the emergence of drug resistance. In particular, the size of the replicating HIV population in vivo remains uncertain. Relatively small population sizes (<1000 infected cells/replication cycle/ infected individual) have been reported, implying that stochastic effects and genetic drift will predominate with the potential for rapid emergence of mutations and shifts in population structure. In contrast, we and others have suggested a relatively large replicating population in vivo, with considerable contribution of deterministic effects, including slow shifts in population structure(45, 47). Most prior studies of HIV diversity in infected patients focused on *env*. *env* datasets are

JVI Accepts published online ahead of print

characterized by high diversity and are rich in strongly selected immune response sites, but do not offer potential to understand detailed emergence of antiretroviral drug resistance. In addition, high genetic diversity presents substantial challenges in obtaining datasets that are not biased by selective amplification; genetic diversity and an excess of insertions and deletions also render *env* datasets difficult to align with confidence, complicating detailed phylogenetic and population genetic analyses.

To investigate HIV population genetics parameters, including population size in regions relevant to antiviral drug resistance to the majority of antiretrovirals, we investigated a 1.3 kb amplicon in *pro-pol*, which includes positions where mutations conferring escape from the CTL response as well as resistance to commonly used treatment regimens are found (33, 51). This region has a degree of genetic diversity (0.8% -2% average pairwise difference (20, 32, 38)) in chronically infected individuals that is suitable for detailed fine structure analyses of HIV populations using phylogenetic (20, 32, 38) and population genetics approaches (1, 3, 11, 32, 45). We investigated *pro-pol* diversity in 33 treatment-naïve individuals by analyzing large collections of individual HIV sequences, in many cases at multiple time intervals after infection. *Pro-pol* diversity varied almost 100-fold, from 0.02% in recently infected individuals, to more than 2% in individuals infected more than 15 years. This new dataset permitted a detailed analysis of HIV genetic variation, from which robust measures diversity, divergence, and population size were obtained. Total sequence diversity (including synonymous and nonsynonymous changes) was strongly correlated with duration of infection even during chronic infection. Increases in genetic diversity over time correlated with increases in synonymous but not nonsynonymous mutations, and did not correlate with plasma HIV RNA level or CD4+ T cell counts. Studies of 11 patients sampled over 1-14 years revealed that the genetic composition of HIV populations changed slowly; significant shifts in HIV populations occurred only after 100-1000 viral generations. We estimated that the effective replicating virus population is at least 10-fold larger than previous measurements derived using analysis of *env* sequences. The size and diversity of the replicating populations suggests that both selection and drift are important mechanisms leading to the emergence of HIV variants *in vivo*.

JVI Accepts published online ahead of print

Materials and Methods

Patients

All HIV infected patients were enrolled in studies of HIV infection at the NIH Clinical Center; patients donated blood samples after giving written informed consent. Duration of infection was estimated in recently infected patients using time of onset of symptoms. All such patients were enrolled in natural history studies of recent HIV infection; all were at least 18 years old, had a recent (<8 weeks) history of an acute febrile illness, consistent with symptomatic HIV infection syndrome following exposure. They also had a history of a nonreactive HIV-1/2 ELISA within a year prior to enrollment or were documented to have plasma HIV >100,000 copies/ml plasma) with an evolving or negative HIV western blot following exposure. Of the remaining patients none had recent history of seroconversion syndrome and the date of the first positive western blot was used to estimate the minimum duration of infection. HIV RNA levels were determined using bDNA Versant version 3.0 (Bayer, Inc.) as previously described (13). CD4 cell subsets were determined by standard clinical immunophenotyping.

Ethics Statement

All participants in this study were enrolled in clinical protocols (00-I-0110, 97-I-0082, 95-I-0072) approved by the NIAID Institutional Review Board (FWA00005897) administered at the NIH Clinical Center in Bethesda, Maryland. Individuals underwent an informed consent process and provided written consent for participation.

Single Genome Sequencing

Plasma from patients was frozen within 5 h of phlebotomy. Specimens were subjected to single genome sequencing as described (23). An amplicon encompassing 297 nt of protease and ca. 700-1200 nt of RT was sequenced. As previously demonstrated (23, 38) *pro-pol* sequences obtained by SGS from each individual patient were highly correlated and were clearly distinguishable from other patient single genome sequence datasets (data not shown). Sequences in Genbank have accession numbers XXX-YYY.

Alignments and Analyses

Sequences were aligned with Clustal W using DNASTAR/Megalign (DNASTAR, Inc; Gap penalty = 2.00, Gap length penalty 2.00). Neighbor joining trees were constructed through Megalign and confirmed in gap stripped neighbor joining trees in

JVI Accepts published online ahead of print

PAUP using pNL4-3 as outgroup; nodes were tested for significance in PAUP using 1000 bootstrap replicates; nodes with >75% bootstrap significance were identified. Measures of diversity (average pairwise distances, denoted APD and expressed as a percent, using p distances to determine pairwise differences; p distance is defined as the number of nucleotide differences between two single genome sequences /total nucleotides sequences (48). In all of these studies, intra-patient p distance determinations were relatively small (<0.03); as described by Nei and Kumar (35) and Nei(34), in the setting of such low p distances, phylogenetic trees using uncorrected p distance provide greater accuracy than trees constructed using more complicated models because of substantial increases in the variance of more complicated models. As expected, therefore, calculating genetic diversity by p distance and Jukes-Cantor corrected p distance yielded nearly identical results that were highly correlated throughout the range of APD (r^2=0.9999).

We obtained an average of 22 (range 9-51) sequences for each time point. To investigate the precision of genetic diversity by this method, we generated model populations with comparable genetic diversity and obtained random samples for genetic diversity determinations.

All polymorphisms (excluding indels) in individual patients were identified and the positions of polymorphisms in each patient alignment were tabulated. Allele frequencies were analyzed with Microsoft Excel based programs.

Replicating population sizes were compared in eleven study patients with longitudinal sampling available. Coalescent estimation of N_e was performed as previously described (58) using the formula:

$\Theta=2N_e\mu$ where Θ is the neutral mutation parameter that defines a neutral coalescence process; for these calculations, Θ is estimated by the nucleotide diversity π, defined in (48) as the average number of nucleotide substitutions per site between two sequences, and μ is the neutral mutation rate per sequence per generation (using 3.4 x 10^{-5} as the per site mutation rate).

Changes in allele frequency were also used to estimate N_e (54, 55, 57) using:

JVI Accepts published online ahead of print

$$F = \frac{1}{n}\sum_{i=1}^{n}\frac{(x_i - y_i)^2}{((1-x_i)+(1-y_i))/2-(x_i y_i)}$$

$$Ne = \frac{t}{2(F - \frac{1}{S1} - \frac{1}{S2})}$$

where n = number of alleles per locus, x_i and y_i represent the allele frequencies at the two time points, t is the number of generations between sampling (1 day/generation), and S1 and S2 are the sample sizes at time 1 and time 2, respectively. Ne was calculated for each site from each patient dataset, and quartile summary statistics were generated.

The geographic subdivision test was carried out as described by Achaz et al. (1). The test statistic, p, is determined by comparing the genetic distance between populations sampled at different times with the distances obtained after repeated shuffling of the same two sets of sequences, and determining how often the distance between the randomized sets exceeded the observed distance. The lower the value of p, the less the chance that two populations arose from the same population (panmixia), with values $p< 1 \text{ x } 10^{-9}$ indicating that population shift has occurred. Statistical tests for significance of correlation coefficient were performed. In analysis of temporally spaced samples, Fisher exact test was used to determine whether differences in allele frequencies between time points were significant. To determine whether an allele was fixed or newly emerged, we studied all positions that were polymorphic at one time point and monomorphic at the other. To identify only those changes that were due to true fixation or emergence, we eliminated those positions in which sampling error could have been responsible for the absence of the minor allele. To eliminate sampling error, we determined the allele frequency at the time the allele was polymorphic, and then calculated the Poisson probability that an allele frequency of zero (not finding polymorphism at that position) at the second time point. For example, if the allele frequency at time 1 = a, then the Possion probability of finding an allele frequency at time 2 =0 is calculated as p(0)= e^{-a} . If p(0) <0.05, then it was statistically unlikely that sampling error was responsible for the absence of polymorphism and we concluded that the polymorphism had arisen or had undergone fixation.

JVI Accepts published online ahead of print

JVI Accepts published online ahead of print

Results

We used single-genome sequencing (SGS, (7, 23, 38)) to study *pro-pol* evolution in 33 HIV-infected treatment naïve patients, all with unprotected sex as their risk category (Table 1). Study patients were predominantly male with an average age of 35.1 years; patients were infected from an estimated 9 days to over 15 years prior to the first sample based on patient history and laboratory studies; 22 patients were infected for <1 year; all but one (patient 1) had a positive western blot at the time of phlebotomy. All patients had CD4 lymphopenia with median 401 CD4 cells/µl blood, and viral RNA levels ranged from 3.1-6.1 $\log_{10}$ copies/ml plasma. As described (23, 38), SGS produces a dataset of individual sequences derived from single HIV genomes that is ideally suited to investigate genetic diversity because of its low error rate, undetectable assay-based recombination, and absence of founder effects due to resampling. We obtained an average of 22 (range 9-51) sequences for each time point. To investigate the precision of these determinations, we constructed theoretical populations, which we sampled with multiple replicates of increasing sample sizes. As shown in Figure 1, increasing sample sizes above 10 sequences yielded adequately precise measurements of genetic diversity (to within 1% of theoretical value, with standard error of mean=0.11). This level of sampling yields reproducible measurements of genetic diversity.

Among these study patients, *pro-pol* nucleotide diversity, as measured by percent average pairwise difference (APD), ranged nearly 100-fold from 0.02% in early (< 1 year duration) infection to slightly more than 2% after 15 years of infection (Table 1). Notably, all but one (patient 1) of the early infection patients had positive western blots, demonstrating that a strong serologic response was already present. We first compared the minimum duration of infection with genetic diversity of each patient sample tested. Overall, there was a significant correlation between the minimum duration of infection and diversity, measured as average pairwise distance (APD) (r^2=0.47, p<0.001), indicating a progressive increase in diversity with time. Detailed analyses revealed that the rate of accumulation of APD was not uniform, however. As shown in Figure 2A, analysis of all samples from the 33 patients revealed that, early in infection (patients 1-13), APD increased relatively rapidly, at an overall rate (0.006 percent/day r^2= 0.45, p=0.002), which approximated that expected from the mutation rate of reverse

JVI Accepts published online ahead of print

transcriptase (corresponding to an increase of 0.004 percent/day, assuming 1 replication cycle/day (31, 39, 40), Figure 2B), and which was similar to previously published data (22). As Keele et. al (24) and others have reported, analysis of HIV genetic diversity in early infections revealed that pairwise differences were Poisson distributed, indicating that overall, mutations occurred randomly throughout the sequence. Consistent with earlier findings, we identified several individuals with recent HIV infection with HIV populations with higher genetic diversity than expected assuming a single infecting virus, indicating infection with more than one founder (Figure 1A).

In contrast to recent infections, when analyses were restricted to the patients infected for more than 1 year, APD increased 0.0002 percent/day (r^2= 0.49, p=0.005 Figure 2C), indicating ongoing accumulation of new mutations, albeit at a rate about 30-fold less than in early infection. During the period where accumulation of diversity slowed (1-2 y), we noted considerable range in diversity among patients (Figure 1A, and Figure 2D), suggesting variable effects of selection and drift.

The period approximating 1 year of infection included samples with a relatively wide spectrum of genetic diversity. To investigate whether mutations were distributed randomly throughout *pro-pol* , we analyzed the distribution of pairwise differences. As previously described, random accumulation of mutations will yield distributions according to Poisson statistics, while nonrandom mutation will result in skewed pairwise differences. Analysis of the distribution of pairwise differences in HIV populations from chronically infected individuals revealed distributions with strong Poisson characteristics, but with deviations from ideal Poisson populations (Maldarelli, unpublished data). These data suggest that mutations continue to accumulate in random fashion during chronic infection, but specific changes may occur as well.

To further characterize the accumulation of new mutations, we compared changes in synonymous and nonsynonymous diversity over time. As shown in Figure 2E, both nonsynonymous and synonymous diversity increased sharply during early months of infection; however, approximately 8 months later, nonsynonymous diversity stabilized and synonymous diversity continued to increase. These data indicate that PR and RT are undergoing change largely under purifying selection most likely as a result of constraints on protein structure.

JVI Accepts published online ahead of print

Although we detected a significant correlation between genetic diversity and duration of infection, the correlation coefficients for recent and chronic infection (r^2=0.45, r^2=0.49, respectively) indicated that duration of infection explained only a portion of the variability in genetic diversity. To look for other correlates, we compared virologic and immunologic measures. No correlation was found between diversity and plasma HIV RNA or CD4 count in individuals with established HIV infection (duration of infection >3 months, Figure 3A and B, r^2=0.04 and r^2=0.07, respectively), indicating that overall HIV *pro-pol* genetic variation was not associated with the level of viremia or extent of immunodeficiency.

We further investigated the relationship between genetic diversity and time with longitudinal sampling of 12 patients with HIV infection and varying baseline diversity. To determine the relative tempo of HIV variation, we compared sequences from samples obtained on a daily, monthly, and yearly basis by phylogenetic analysis. As we and others have shown (22, 24, 41, 49, 59), HIV population structure was relatively monomorphic during early infection (Figure 4, patient 2, panel D) , which arose from the few mutations that appeared over the relatively short period of observation. Early in infection (<1 y), diversity increased approximately as predicted by the mutation rate (Figure 4, patient 8) as previously noted (22). By contrast, during chronic HIV infection, diversity remained relatively stable (Figure 4, patients 11, 19, 24,25,26 panel B), even during progressive decline of CD4 cell counts (Patients 19 and 25) and more than 10-fold increases in HIV RNA levels (Patient 25). As expected from cross sectional data (Figure 2A), increases in diversity were, nonetheless, detectable in temporally spaced samples obtained from individual patients, although consistent rates of divergence among all patients were, in general, not discernible (data not shown). Analysis of daily samples from two patients revealed no variation in HIV diversity over a 10 day observation period (Figure 4, patient 24, daily samples, panel B, second patient not shown), excluding rapid fluctuation due, for example, to differential seeding of the population from genetically distinct tissue sites of replication.

Neighbor-joining analysis revealed that temporally spaced *pro-pol* sequences remained highly related to one another. Samples obtained on a daily basis (Figure 4, patient 24) or over 5 years revealed that only a few (1-6) sequences or clusters of

JVI Accepts published online ahead of print

sequences from individual times had bootstrap values (>75%) sufficient to support the observed branching (Figure 4, patients 8, 11, 24, 25, 26 panels D, thick colored bars). Of the 12 patients with longitudinal sampling, one (Figure 4, patient 25, panel D, blue branches) had evidence for divergence in a subset of 6 sequences after a sampling interval exceeding five years,and a second (patient 19) had evidence of emergence of a distint lineage after nearly 14 years. In the remaining patients, phylogenetic topologies of temporally spaced samples suggested a shared common ancestry for HIV sequences; the most recent sequences did not demonstrate progressive accumulation of diversity compared to the earliest sequences.

Temporally spaced data were also useful in providing a detailed view of HIV polymorphisms and identify changes in individual allele frequencies over time. As shown in Table 2, for 10/11 patients a relatively small number of alleles underwent change during the observation period (median 9%, range 0-18%). None of the alleles that emerged or underwent fixation were linked to alleles that underwent significant change in allele frequency, indicating that fixation did not result in a selective sweep that carried other alleles. Rather, the occurrence of unlinked polymorphisms emerging or undergoing fixation in this fashion indicates that populations are highly diverse, and certain lineages were simply lost or emerged as result of new mutation. Most sites did not undergo changes in allele frequency, suggesting that selection at these sites was not sufficiently strong to change the frequency. HIV from one patient (patient 19) underwent significant change during a prolonged observation period (5099 d) with 43% of 90 polymorphisms undergoing significant change, 21 of which were new or lost alleles, a number of which were linked (Table 2). As shown in Figure 4, the HIV population structure in this patient was distinct, with all of the sequences from the later time point on a distinct lineage with strong bootstrap support, accounting for the number of new changes. Patient 25 also had a new bootstrap supported lineage emerge after a long period (5.7 years), but also had a number of variants present.

Recombination is a common phenomenon in HIV replication; as we previously reported, approximately 6% of infected cells are likely infected with more than one provirus (21), providing the opportunity for recombination to occur. In HIV infected

JVI Accepts published online ahead of print

patients, recombinants accrue during the entire course of infection. As a result, demonstration of recombination using standard phylogenetic techniques (18) detected frequent evidence of recombination with recombination intervals of 36-120 nt (Maldarelli, unpublished observation).

Despite the absence of clear phylogenetic evidence of divergence and the relatively stable intrapatient viral diversity, substantial population shifts were detectable when we applied an adaptation of the geographic subdivision test (1) to identify patterns of population structure. Population shift is indicated by a loss of panmixia, (a population characteristic in which all sequences in the sample comparison belong to a single replicating group); in comparing sequences from two different time points, a low (1 X 10^{-9}) probability of panmixia indicates population divergence. In contrast to the relatively homogeneous populations indicated by the NJ analyses, the geographic subdivision test showed clear evidence of population shift in *pro-pol* sequences from patients with HIV infection sampled over prolonged periods (Figure 4, panels C, patients 8, 11, 19, 24,25,26), whereas at short intervals (patient 2 or patient 24, daily samples) no evidence of population shift was detectable. Cumulative analysis of all intrapatient pairwise comparisons revealed that the median time to population shift (defined as a probability of panmixia $<10^{-9}$) was 1017 days, and the minimum duration before shift was detected was 193 days (Figure 5). These data are consistent with our initial report of the population subdivision adaptation (1) and indicate that significant change in HIV *pro-pol* population structure takes place with a time scale that is 100-1000 fold longer than the replication cycle time of HIV *in vivo* (1-2 d).

The relatively slow rate of population shift in HIV population structure implies relatively large replicating populations *in vivo*. Therefore, we used two tests to investigate further the effective size (N_e) of the HIV populations. As shown in Figure 6, coalescent analyses (diamonds) yielded uniformly low measures of effective population size, on the order of 100 to 1000, similar to estimates previously reported (7, 10, 14, 50, 53), a surprising result in light of the slow change in population structure detected by the population subdivision analyses. This difference may be due to the fact that this method ignores the contribution of selection and recombination, both of which can lead to underestimation of population size (29, 45). Therefore, we next determined population

JVI Accepts published online ahead of print

size using a phylogeny-independent method described by Nei and Tajima (55) and Waples (57). This method estimates population size based on the rate of change of individual allele frequencies over time and thus yields a range of population sizes; assuming no selection, large changes in allele frequencies yield the smallest estimates of population size, and relatively small changes in allele frequency yield the largest population sizes.As shown in Figure 6A (Whisker plot), N_e estimates varied by more than 10-100-fold among individual patients, reflecting a wide range of changes in allele frequency among HIV *pro-pol* polymorphisms. HIV populations from two individuals (patients 10, 14) had relatively narrow quartile distributions of population sites, reflecting restricted range of allele frequency changes.

The median N_e estimates obtained using the latter method were in the range of 10^3-10^4 (Figure 3B), or >30-fold higher than that measured by coalescent-based methods, and are more consistent with, although still less than, population sizes estimated from linkage equilibrium analyses (45). Even the minimum estimates of allele frequencies obtained by this method were, in general, greater than those estimated by coalescent methods. To investigate the relative contributions of selection and drift on N_e, we further analyzed the type of variability on a site by site basis (Figure 6B). We expected that nonsynonymous polymorphisms resulting in changes in amino acids would be subject to greater selective forces and would yield smaller values for N_e, whereas estimates of N_e using synonymous polymorphisms would be less subject to selection and more influenced by genetic drift, and would yield large N_e. Consistent with this expectation, the overall population size measured using synonymous sites was greater than that measured using nonsynonymous sites; the difference, however, was modest and of marginal statistical significance (5,600 vs. 4,500 transmitting cells per generation for nonsynonymous and synonymous sites, respectively, two sided t test, $p = 0.035$). We investigated the estimates of population sizes by nucleotide position of polymorphisms within *pro-pol* to investigate the role of synonymous and nonsynonymous sites and to determine whether there were region-specific effects of drift or selection. As shown in Figure 6B, the nonsynonymous and synonymous alleles that contributed to large and small population size estimates were distributed throughout *pro-pol* and were not localized by gene or domain.

JVI Accepts published online ahead of print

The observation that some nonsynonymous sites yielded high population sizes suggests that some sites are not undergoing selection; alternatively, it is possible that such polymorphisms are maintained by frequent mutation at specific sites. If individual sites were undergoing frequent mutation, we would expect to identify such sites as repeatedly polymorphic in several individuals. However, of 56 nonsynonymous sites yielding population estimates >20,000, only 2 (3.6%) were present more than once. As a result, it is unlikely that frequent mutation at individual sites is responsible for persistence of stable polymorphisms; these polymorphisms are more likely to be stably maintained over time because of relatively large population sizes. Taken together, these data indicate that measurements of HIV effective population sizes are heavily influenced by variations in allele frequency and change over time and from one site to the next due to variation in selection and drift. Our estimates should, therefore, be taken as a lower bound, and the true values are likely to be much higher.

JVI Accepts published online ahead of print

Discussion

HIV genetic diversity within individuals is the substrate upon which immune and antiretroviral drug selection act. Previous studies (22, 24, 37, 41, 49, 59) have reported that diversity in most recently infected individuals is very low, consistent with initiation of infection with a single variant. In patients with established infection, *pro-pol* diversity accumulated at a much lower rate than in recently infected individuals, and over the course of infection, diversity increased in a non-linear fashion (Figure 2A). The strength of the correlation between diversity and time for both early and established HIV infection (r^2=0.47-0.55) suggests that duration of infection only explains a portion of the variability in diversity. All of the participants in this study were infected with subtype B virus; a recent study sequencing single genomes from early post-infection subtype C infected individuals has identified a similar increase in genetic diversity in nonstructural genes including *vif*, *vpu*, *tat* and *rev* (43).

The absence of association between *pro-pol* diversity and viral RNA level that we observed is similar to a previous analysis of *env* diversity and viral RNA levels (4), and implies that, despite 100 to 1,000-fold differences in the level of viremia, the number of productively infected cells must be sufficiently large to sustain a highly diverse population of virus. Furthermore, we found no instances of a sudden shift in the HIV population that would suggest a bottleneck due to a selective sweep or other strong limitation on the infected cell population size. Additionally, the absence of short term fluctuations in diversity implies that the virus in blood is a well mixed population derived from a constant, steady source, rather than localized bursts of virus from sites infected with genetically distinct populations. Finally, in a related study, we have found that diversity of the virus population is maintained throughout the course of infection, even following reductions in the number of productively infected cells by 10,000 fold following antiretroviral therapy, indicating a large population of infected cells (Kearney, et al., presented at the 17th Conference on Retroviruses and Opportunistic Infections, San Francisco CA, Feb 16-19, 2010). As previously observed (22), genetic diversity early in HIV infection accumulated at a rate approximating that expected from its mutation rate. In contrast, accumulation of diversity slowed by more than 30-fold in chronically infected individuals, suggesting a restriction on accumulation of new mutations.

JVI Accepts published online ahead of print

Differential accumulation of synonymous and nonsynonymous mutations is consistent with limitation of diversity due to purifying selection. In general only a small proportion of polymorphisms underwent change over time, fewer still were fixed and only in one patient (patient 19), with strong phylogenetic evidence of emergence of a distinct variant more than 13 years after infection, were these fixed polymorphisms linked (Table 2) . Previous reports of accumulation of variation in *env* according to a strict (51) or relaxed (25) molecular clock were not reflected in our overall analysis of *pro-pol*. Instead, diversity increased asymptotically, with maximum APD values on the order of 2% about 15 years after infection, suggesting a limit to the amount of diversity that can accumulate within an individual. Similar conclusions (45) on the lack of temporal structure in HIV sequences have been drawn from analyses of *env* sequences in several patients (5). Maximum intrapatient *pro-pol* diversity during chronic infection was still substantially lower than the corresponding interpatient pairwise comparisons, which typically exceeded 5% ((22) and data not shown). In addition it is not clear why accumulation of diversity slowed markedly after 9-18 months of infection. It is unlikely that slowing in diversity accumulation was the result of onset of immune responses, as accumulation of diversity occurred after development of serologic and cellular immune responses. These data indicate that, within an individual, HIV genetic variation remains restricted, by strong purifying selective forces.

All of the participants in this study were infected with subtype B virus. It will be of great interest to determine whether other subtypes have similar intrapatient diversity, and accumulate diversity at rates comparable to subtype B. Recently, Rossenkhan and coworkers (43) conducted a detailed analysis of subtype C infected individuals, sequencing single genomes from early post-infection individuals to obtain diversity estimates for HIV accessory genes including *vif*, *vpu*, *tat* and *rev*. Similar to subtype B, genetic diversity was restricted in these early infection samples and accumulated over time. A comprehensive analysis of subtype specific genetic variation will yield new insights in understanding HIV pathogenesis.

The relative size of the replicating HIV population (N_e) remains uncertain, but is a critical parameter in understanding the spread of new mutations conferring resistance and immune escape (8, 9, 37). In relatively small populations (<<1/mutation rate or

JVI Accepts published online ahead of print

<<3X10^4), new mutations spread in stochastic fashion, while in large populations (>>1/mutation rate or >>3X10^4), emergence of new variants approaches a deterministic limit (47). Estimating replicating population sizes typically uses coalescent approaches. Coalescent theory is an inherently retrospective approach rooted in neutral population genetics theory that reconstructs a genetic history based on present population structure. The model assumes mutations arise according to a constant mutation rate in a strict molecular clock-like fashion; all alleles are neutral, reassort in random mating in populations that remain constant in size. Using a contemporaneous set of polymorphisms with measured allele frequencies in populations, coalescence uses probability analyses to reconstruct an entire population history, identifies times when genealogies "coalesce" to a most recent common ancestor (MRCA) of the population, and describes the most probable pathway to the ancestor, depicted in dendrograms that are measured in time (rather that genetic distances present in phylogenetic analyses). Based on genetic diversity determinations, a replicating population size can be estimated. Coalescence theory generally underestimates population size, but represents a powerful approach to reconstructing genetic histories of diverse variants including HIV (60) over long periods, where genetic diversity is substantial. In analysis of intra patient data, however, the genetic diversity is more restricted, and coalescent approaches may be more sensitive to the effects of selection, yielding lower estimates for population size. In our estimates, standard coalescent approaches yielded uniformly low replicating population sizes, in the range of 10 to 100 (Figure 3). Additional analyses using allele frequency variation to estimate N_e yielded replicating population sizes that were 30- fold greater than by coalescent based estimates, and these estimates varied greatly from one site to the next.

Site by site analysis also revealed that both synonymous and nonsynonymous polymorphisms underwent relatively slow change, indicating that some nonsynonymous sites are subject to relatively little selection. In addition, we also observed nonsynonymous and synonymous sites that underwent change at relatively rapid rate, suggesting that such sites were undergoing selection compared to others. Constraints on nonsynonymous sites have been well described: additional selective forces, including RNA structure and codon preference, may affect the allele frequency of synonymous sites. One consequence of large population sizes is a relatively long time to detectable

JVI Accepts published online ahead of print

genetic shift. The median time of approximately 1000 days (corresponding to about 1000 virus generations) for population genetic shift to appear suggests that, prior to therapy, HIV replication proceeds as a large, well mixed population without selective sweeps or rapid changes in composition. Since many, if not most, of the nonsynonymous changes in HIV that become fixed during all phases of infection are in sites recognized by the cellular or humoral immune response (22, 26, 28, 56), the absence of detectable bottlenecks in the population associated with their appearance implies that the selective force imposed by the immune response to any given epitope, although readily detectable by the selection of escape mutations, is not sufficiently strong to influence the overall population size or structure.

Our finding of relatively large population sizes contrasts sharply with previous studies that concluded the existence of relatively small population sizes using *env* sequences for analyses. Earlier *env* datasets available for study, such as Shankarappa (51) are extensive, but have relatively few individual plasma-derived sequences compared to the larger numbers of sequences used here to determine population size. For comparison purposes, we did carry out a site by site analysis on two patients in the Shankarappa dataset with 10-11 sequences/time point. Our analysis revealed median population sizes of 2736 (range 2362-53702) and 5688 (range 3197-62571) similar to what we have identified in *pro-pol*; the high upper boundaries of these determinations represent the contribution of alleles with relatively stable allele frequency over time and reflect the presence of relatively large population size. New studies with more single genome sequences will be useful in directly estimating population sizes using *env* and *pro-pol* sequences.

Population sizes in the range of 1 x 10^4 to 1 x 10^5 approximate the inverse of the estimated unselected mutation rate of 3-4 X10^{-5} (31); HIV mutation rate in vivo has not been well studied, and actual mutation frequencies are likely to be strongly influenced by both selection and genetic drift (12, 14, 44-46). This conclusion is consistent with the detection of alleles with rapid (selection) and slow (drift) change and with the overall slowing in accumulation of diversity in chronic HIV infection. The issue becomes particularly important in considering the frequency of drug resistance mutations in untreated individuals. The rapid and reproducible appearance of such mutations

JVI Accepts published online ahead of print

following monotherapy with antiviral drugs such as 3TC (16), and single-dose nevirapine (19) implies their presence in the replicating virus population in most or all infected individuals prior to therapy. Their frequency will be determined by the balance between mutation, counterselection, and drift (47), but must be at least the inverse of the replicating population size, on average. Studies to date using sensitive allele-specific PCR methods, however, have failed to reproducibly detect such mutations, suggesting that the population size may be substantially larger than estimated here. Further development of very sensitive mutation detection technology as well as advances in mathematical modeling will be needed to resolve this important issue and provide critical tests of the selection-drift hypothesis and a better understanding of the virus population size and structure, which can be directly applied to understanding the emergence of drug resistance.

The population studies reported here have broad implications for understanding the pathogenesis and therapeutic responses in other chronic viral infections, especially so for those viruses with constantly replicating populations in chronic infection and new and expanding therapeutic agents, such as hepatitis B and C. Hepatitis B has a number of effective therapeutic agents, although determinants of viral control and resistance are poorly understood. Genetic diversity is substantial, but the relationships between genetic diversity, population size and emergence of resistance have not been extensively investigated (52, 61). Therapy for hepatitis C has expanded with additional targets and therapeutic agents; cure rates have improved but the virologic correlates of eradication are incompletely understood. Population genetics studies have demonstrated hepatitis C populations are highly genetically diverse, even relative to HIV, so it is likely that, similar to HIV, drug resistant mutations will pre-exist prior to therapy. Intrapatient genetic variation has been investigated (6, 17, 27, 30, 36), although population sizes have not been extensively investigated and it is not known how fast new drug resistant mutations may be expected to emerge. Additional studies, such as those reported here will have direct applications in the design of clinical trials and the composition of combination therapy necessary to eradicate viral infection.

JVI Accepts published online ahead of print

Acknowledgments

We thank Marguerite Van Houtte, Brendan Larder, Werner Verbiest, Wei-Shau Hu, Vinay Pathak, and Justin Palmer for discussions and technical assistance. We are indebted to H. Masur and H.C. Lane for support and insightful discussions. JMC was a research Professor of the American Cancer Society with support from the F.M. Kirby Foundation.

JVI Accepts published online ahead of print

JVI Accepts published online ahead of print

References

1. **Achaz, G., S. Palmer, M. Kearney, F. Maldarelli, J. W. Mellors, J. M. Coffin, and J. Wakeley.** 2004. A robust measure of HIV-1 population turnover within chronically infected individuals. Mol Biol Evol **21:**1902-12.
2. **Althaus, C. L., and S. Bonhoeffer.** 2005. Stochastic interplay between mutation and recombination during the acquisition of drug resistance mutations in human immunodeficiency virus type 1. J Virol **79:**13572-8.
3. **Batorsky, R., M. F. Kearney, S. E. Palmer, F. Maldarelli, I. M. Rouzine, and J. M. Coffin.** Estimate of effective recombination rate and average selection coefficient for HIV in chronic infection. Proc Natl Acad Sci U S A **108:**5661-6.
4. **Bello, G., C. Casado, S. Garcia, C. Rodriguez, J. del Romero, A. V. Borderia, and C. Lopez-Galindez.** 2004. Plasma RNA viral load is not associated with intrapatient quasispecies heterogeneity in HIV-1 infection. Arch Virol **149:**1761-71.
5. **Bello, G., C. Casado, S. Garcia, C. Rodriguez, J. del Romero, A. Carvajal-Rodriguez, D. Posada, and C. Lopez-Galindez.** 2007. Lack of temporal structure in the short term HIV-1 evolution within asymptomatic naive patients. Virology **362:**294-303.
6. **Bernini, F., E. Ebranati, C. De Maddalena, R. Shkjezi, L. Milazzo, A. Lo Presti, M. Ciccozzi, M. Galli, and G. Zehender.** Within-host dynamics of the hepatitis C virus quasispecies population in HIV-1/HCV coinfected patients. PLoS One **6:**e16551.
7. **Brown, A. J.** 1997. Analysis of HIV-1 env gene sequences reveals evidence for a low effective number in the viral population. Proc Natl Acad Sci U S A **94:**1862-5.
8. **Coffin, J. M.** 1995. HIV population dynamics in vivo: implications for genetic variation, pathogenesis, and therapy. Science **267:**483-9.
9. **Daar, E. S., and D. D. Richman.** 2005. Confronting the emergence of drug-resistant HIV type 1: impact of antiretroviral therapy on individual and population resistance. AIDS Res Hum Retroviruses **21:**343-57.
10. **Drummond, A. J., G. K. Nicholls, A. G. Rodrigo, and W. Solomon.** 2002. Estimating mutation parameters, population history and genealogy simultaneously from temporally spaced sequence data. Genetics **161:**1307-20.
11. **Dykes, C., J. Najjar, R. J. Bosch, M. Wantman, M. Furtado, S. Hart, S. M. Hammer, and L. M. Demeter.** 2004. Detection of drug-resistant minority variants of HIV-1 during virologic failure of indinavir, lamivudine, and zidovudine. J Infect Dis **189:**1091-6.
12. **Edwards, C. T., E. C. Holmes, O. G. Pybus, D. J. Wilson, R. P. Viscidi, E. J. Abrams, R. E. Phillips, and A. J. Drummond.** 2006. Evolution of the human immunodeficiency virus envelope gene is dominated by purifying selection. Genetics **174:**1441-53.
13. **Elbeik, T., W. G. Alvord, R. Trichavaroj, M. de Souza, R. Dewar, A. Brown, D. Chernoff, N. L. Michael, P. Nassos, K. Hadley, and V. L. Ng.** 2002. Comparative analysis of HIV-1 viral load assays on subtype quantification: Bayer Versant HIV-1 RNA 3.0 versus Roche Amplicor HIV-1 Monitor version 1.5. J Acquir Immune Defic Syndr **29:**330-9.

JVI Accepts published online ahead of print

14. **Frost, S. D., M. J. Dumaurier, S. Wain-Hobson, and A. J. Brown.** 2001. Genetic drift and within-host metapopulation dynamics of HIV-1 infection. Proc Natl Acad Sci U S A **98:**6975-80.

15. **Frost, S. D., and A. R. McLean.** 1994. Quasispecies dynamics and the emergence of drug resistance during zidovudine therapy of HIV infection. AIDS **8:**323-32.

16. **Frost, S. D., M. Nijhuis, R. Schuurman, C. A. Boucher, and A. J. Brown.** 2000. Evolution of lamivudine resistance in human immunodeficiency virus type 1-infected individuals: the relative roles of drift and selection. J Virol **74:**6262-8.

17. **Honda, M., S. Kaneko, A. Sakai, M. Unoura, S. Murakami, and K. Kobayashi.** 1994. Degree of diversity of hepatitis C virus quasispecies and progression of liver disease. Hepatology **20:**1144-51.

18. **Hudson, R. R.** 1983. Properties of a neutral allele model with intragenic recombination. Theor Popul Biol **23:**183-201.

19. **Johnson, J. A., J. F. Li, L. Morris, N. Martinson, G. Gray, J. McIntyre, and W. Heneine.** 2005. Emergence of drug-resistant HIV-1 after intrapartum administration of single-dose nevirapine is substantially underestimated. J Infect Dis **192:**16-23.

20. **Jordan, M. R., M. Kearney, S. Palmer, W. Shao, F. Maldarelli, E. P. Coakley, C. Chappey, C. Wanke, and J. M. Coffin.** Comparison of standard PCR/cloning to single genome sequencing for analysis of HIV-1 populations. J Virol Methods **168:**114-20.

21. **Josefsson, L., M. S. King, B. Makitalo, J. Brannstrom, W. Shao, F. Maldarelli, M. F. Kearney, W. S. Hu, J. Chen, H. Gaines, J. W. Mellors, J. Albert, J. M. Coffin, and S. E. Palmer.** Majority of CD4+ T cells from peripheral blood of HIV-1-infected individuals contain only one HIV DNA molecule. Proc Natl Acad Sci U S A **108:**11199-204.

22. **Kearney, M., F. Maldarelli, W. Shao, J. B. Margolick, E. S. Daar, J. W. Mellors, V. Rao, J. M. Coffin, and S. Palmer.** 2009. Human immunodeficiency virus type 1 population genetics and adaptation in newly infected individuals. J Virol **83:**2715-27.

23. **Kearney, M., S. Palmer, F. Maldarelli, W. Shao, M. A. Polis, J. Mican, D. Rock-Kress, J. B. Margolick, J. M. Coffin, and J. W. Mellors.** 2008. Frequent polymorphism at drug resistance sites in HIV-1 protease and reverse transcriptase. AIDS **22:**497-501.

24. **Keele, B. F., E. E. Giorgi, J. F. Salazar-Gonzalez, J. M. Decker, K. T. Pham, M. G. Salazar, C. Sun, T. Grayson, S. Wang, H. Li, X. Wei, C. Jiang, J. L. Kirchherr, F. Gao, J. A. Anderson, L. H. Ping, R. Swanstrom, G. D. Tomaras, W. A. Blattner, P. A. Goepfert, J. M. Kilby, M. S. Saag, E. L. Delwart, M. P. Busch, M. S. Cohen, D. C. Montefiori, B. F. Haynes, B. Gaschen, G. S. Athreya, H. Y. Lee, N. Wood, C. Seoighe, A. S. Perelson, T. Bhattacharya, B. T. Korber, B. H. Hahn, and G. M. Shaw.** 2008. Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection. Proc Natl Acad Sci U S A **105:**7552-7.

25. **Lemey, P., S. L. Kosakovsky Pond, A. J. Drummond, O. G. Pybus, B. Shapiro, H. Barroso, N. Taveira, and A. Rambaut.** 2007. Synonymous

JVI Accepts published online ahead of print

substitution rates predict HIV disease progression as a result of underlying replication dynamics. PLoS Comput Biol **3:**e29.

26. **Leslie, A. J., K. J. Pfafferott, P. Chetty, R. Draenert, M. M. Addo, M. Feeney, Y. Tang, E. C. Holmes, T. Allen, J. G. Prado, M. Altfeld, C. Brander, C. Dixon, D. Ramduth, P. Jeena, S. A. Thomas, A. St John, T. A. Roach, B. Kupfer, G. Luzzi, A. Edwards, G. Taylor, H. Lyall, G. Tudor-Williams, V. Novelli, J. Martinez-Picado, P. Kiepiela, B. D. Walker, and P. J. Goulder.** 2004. HIV evolution: CTL escape mutation and reversion after transmission. Nat Med **10:**282-9.

27. **Liu, L., B. E. Fisher, K. A. Dowd, J. Astemborski, A. L. Cox, and S. C. Ray.** Acceleration of hepatitis C virus envelope evolution in humans is consistent with progressive humoral immune selection during the transition from acute to chronic infection. J Virol **84:**5067-77.

28. **Liu, Y., J. P. McNevin, S. Holte, M. J. McElrath, and J. I. Mullins.** Dynamics of viral evolution and CTL responses in HIV-1 infection. PLoS One **6:**e15639.

29. **Liu, Y., and J. E. Mittler.** 2008. Selection dramatically reduces effective population size in HIV-1 infection. BMC Evol Biol **8:**133.

30. **Liu, Z., D. M. Netski, Q. Mao, O. Laeyendecker, J. R. Ticehurst, X. H. Wang, D. L. Thomas, and S. C. Ray.** 2004. Accurate representation of the hepatitis C virus quasispecies in 5.2-kilobase amplicons. J Clin Microbiol **42:**4223-9.

31. **Mansky, L. M., and H. M. Temin.** 1995. Lower in vivo mutation rate of human immunodeficiency virus type 1 than that predicted from the fidelity of purified reverse transcriptase. J Virol **69:**5087-94.

32. **Mens, H., M. Kearney, A. Wiegand, W. Shao, K. Schonning, J. Gerstoft, N. Obel, F. Maldarelli, J. W. Mellors, T. Benfield, and J. M. Coffin.** HIV-1 continues to replicate and evolve in patients with natural control of HIV infection. J Virol **84:**12971-81.

33. **Mullins, J. I., and M. A. Jensen.** 2006. Evolutionary dynamics of HIV-1 and the control of AIDS. Curr Top Microbiol Immunol **299:**171-92.

34. **Nei, M.** 1987. Molecular Evolutionary Genetics. Columbia University Press, New York.

35. **Nei, M., Kumar, S.** 2000. Molecular Evolution and Phylogenetics. Oxford University Press, Inc, New York

36. **Netski, D. M., Q. Mao, S. C. Ray, and R. S. Klein.** 2008. Genetic divergence of hepatitis C virus: the role of HIV-related immunosuppression. J Acquir Immune Defic Syndr **49:**136-41.

37. **Overbaugh, J., and C. R. Bangham.** 2001. Selection forces and constraints on retroviral sequence variation. Science **292:**1106-9.

38. **Palmer, S., M. Kearney, F. Maldarelli, E. K. Halvas, C. J. Bixby, H. Bazmi, D. Rock, J. Falloon, R. T. Davey, Jr., R. L. Dewar, J. A. Metcalf, S. Hammer, J. W. Mellors, and J. M. Coffin.** 2005. Multiple, linked human immunodeficiency virus type 1 drug resistance mutations in treatment-experienced patients are missed by standard genotype analysis. J Clin Microbiol **43:**406-13.

JVI Accepts published online ahead of print

39. **Perelson, A. S., P. Essunger, Y. Cao, M. Vesanen, A. Hurley, K. Saksela, M. Markowitz, and D. D. Ho.** 1997. Decay characteristics of HIV-1-infected compartments during combination therapy. Nature **387:**188-91.

40. **Perelson, A. S., A. U. Neumann, M. Markowitz, J. M. Leonard, and D. D. Ho.** 1996. HIV-1 dynamics in vivo: virion clearance rate, infected cell life-span, and viral generation time. Science **271:**1582-6.

41. **Poss, M., H. L. Martin, J. K. Kreiss, L. Granville, B. Chohan, P. Nyange, K. Mandaliya, and J. Overbaugh.** 1995. Diversity in virus populations from genital secretions and peripheral blood from women recently infected with human immunodeficiency virus type 1. J Virol **69:**8118-22.

42. **Rong, L., M. A. Gilchrist, Z. Feng, and A. S. Perelson.** 2007. Modeling within-host HIV-1 dynamics and the evolution of drug resistance: trade-offs between viral enzyme function and drug susceptibility. J Theor Biol **247:**804-18.

43. **Rossenkhan, R., V. Novitsky, T. K. Sebunya, R. Musonda, B. A. Gashe, and M. Essex.** Viral diversity and diversification of major non-structural genes vif, vpr, vpu, tat exon 1 and rev exon 1 during primary HIV-1 subtype C infection. PLoS One **7:**e35491.

44. **Rouzine, I. M., and J. M. Coffin.** 2005. Evolution of human immunodeficiency virus under selection and weak recombination. Genetics **170:**7-18.

45. **Rouzine, I. M., and J. M. Coffin.** 1999. Linkage disequilibrium test implies a large effective population number for HIV in vivo. Proc Natl Acad Sci U S A **96:**10758-63.

46. **Rouzine, I. M., and J. M. Coffin.** 1999. Search for the mechanism of genetic variation in the pro gene of human immunodeficiency virus. J Virol **73:**8167-78.

47. **Rouzine, I. M., A. Rodrigo, and J. M. Coffin.** 2001. Transition between stochastic evolution and deterministic evolution in the presence of selection: general theory and application to virology. Microbiol Mol Biol Rev **65:**151-85.

48. **Rozas, J., J. C. Sanchez-DelBarrio, X. Messeguer, and R. Rozas.** 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. Bioinformatics **19:**2496-7.

49. **Salazar-Gonzalez, J. F., E. Bailes, K. T. Pham, M. G. Salazar, M. B. Guffey, B. F. Keele, C. A. Derdeyn, P. Farmer, E. Hunter, S. Allen, O. Manigart, J. Mulenga, J. A. Anderson, R. Swanstrom, B. F. Haynes, G. S. Athreya, B. T. Korber, P. M. Sharp, G. M. Shaw, and B. H. Hahn.** 2008. Deciphering human immunodeficiency virus type 1 transmission and early envelope diversification by single-genome amplification and sequencing. J Virol **82:**3952-70.

50. **Seo, T. K., J. L. Thorne, M. Hasegawa, and H. Kishino.** 2002. Estimation of effective population size of HIV-1 within a host: a pseudomaximum-likelihood approach. Genetics **160:**1283-93.

51. **Shankarappa, R., J. B. Margolick, S. J. Gange, A. G. Rodrigo, D. Upchurch, H. Farzadegan, P. Gupta, C. R. Rinaldo, G. H. Learn, X. He, X. L. Huang, and J. I. Mullins.** 1999. Consistent viral evolutionary changes associated with the progression of human immunodeficiency virus type 1 infection. J Virol **73:**10489-502.

52. **Sheldon, J., B. Ramos, J. Garcia-Samaniego, P. Rios, A. Bartholomeusz, M. Romero, S. Locarnini, F. Zoulim, and V. Soriano.** 2007. Selection of hepatitis

JVI Accepts published online ahead of print

B virus (HBV) vaccine escape mutants in HBV-infected and HBV/HIV-coinfected patients failing antiretroviral drugs with anti-HBV activity. J Acquir Immune Defic Syndr **46:**279-82.

53. **Shriner, D., R. Shankarappa, M. A. Jensen, D. C. Nickle, J. E. Mittler, J. B. Margolick, and J. I. Mullins.** 2004. Influence of random genetic drift on human immunodeficiency virus type 1 env evolution during chronic infection. Genetics **166:**1155-64.

54. **Tajima, F.** 1983. Evolutionary relationship of DNA sequences in finite populations. Genetics **105:**437-60.

55. **Tajima, F., and M. Nei.** 1984. Note on genetic drift and estimation of effective population size. Genetics **106:**569-74.

56. **Troyer, R. M., J. McNevin, Y. Liu, S. C. Zhang, R. W. Krizan, A. Abraha, D. M. Tebit, H. Zhao, S. Avila, M. A. Lobritz, M. J. McElrath, S. Le Gall, J. I. Mullins, and E. J. Arts.** 2009. Variable fitness impact of HIV-1 escape mutations to cytotoxic T lymphocyte (CTL) response. PLoS Pathog **5:**e1000365.

57. **Waples, R. S.** 1989. A generalized approach for estimating effective population size from temporal changes in allele frequency. Genetics **121:**379-91.

58. **Watterson, G. A.** 1975. On the number of segregating sites in genetical models without recombination. Theor Popul Biol **7:**256-76.

59. **Wolinsky, S. M., C. M. Wike, B. T. Korber, C. Hutto, W. P. Parks, L. L. Rosenblum, K. J. Kunstman, M. R. Furtado, and J. L. Munoz.** 1992. Selective transmission of human immunodeficiency virus type-1 variants from mothers to infants. Science **255:**1134-7.

60. **Yusim, K., M. Peeters, O. G. Pybus, T. Bhattacharya, E. Delaporte, C. Mulanga, M. Muldoon, J. Theiler, and B. Korber.** 2001. Using human immunodeficiency virus type 1 sequences to infer historical features of the acquired immune deficiency syndrome epidemic and human immunodeficiency virus evolution. Philos Trans R Soc Lond B Biol Sci **356:**855-66.

61. **Zoulim, F., and S. Locarnini.** 2009. Hepatitis B virus resistance to nucleos(t)ide analogues. Gastroenterology **137:**1593-608 e1-2.

JVI Accepts published online ahead of print

Figure Legends

Figure 1.

Determining precision of SGS. (A) Theoretical Poisson-distributed populations of 1000 sequences with average pairwise difference of 1% were generated. Seven replicate samples of increasing numbers of sequences from 2-100 sequences per sample were obtained and APD determined .(B) Standard deviation of the APD determinations.

Figure 2.

Non-linear accumulation of HIV diversity over time.

A, Overall diversity expressed as percent average pairwise difference was determined from alignments of *pro-pol* sequences obtained from all samples from patients 1-27 and presented as a function of minimum duration of infection as defined in Materials and Methods. Patients for whom only a single sample was available for analysis are shown in black. B. Accumulation of mutations in recently infected individuals (patients 1-13). C. Accumulation of mutations in chronically infected individuals (patients 14-27). D. Accumulation of mutations during 0.5-3 y duration.

E. Diversity measurements were obtained separately for synonymous (red squares) and nonsynonymous (blue triangles) sites from SGS datasets using DNASP, and are presented as a function of time after infection. To avoid overweighting of patients with multiple samples, only the earliest time point for each patient was included for analyses in B-E.

Figure 3.

No correlation between HIV genetic diversity and level of viremia or CD4 cell concentration Overall diversity expressed as percent average pairwise difference was determined from alignments of *pro-pol* sequences obtained from all samples from patients 1-27 and presented as a function of minimum duration of infection as defined in Materials and Methods. Correlation between diversity and viral RNA level (A) or CD4+ T cell count (B). Only the earliest time point from each patient was included for analysis.

JVI Accepts published online ahead of print

Figure 4.

HIV *pro-po*l diversity and population shifts in HIV infected patients.

Each patient enrolled in the study underwent phlebotomy at the study days indicated. A. The level of viremia (boxes) and CD4 lymphopenia (diamonds) was determined. Samples indicated by colored circles were subjected to SGS. B. Sequences obtained by SGS at the indicated times were aligned and APD was determined. C. Sequences from the indicated time points were compared to the sequence set from the earliest time point in the patient dataset, and the probability of panmixia was calculated (1). D. Neighbor joining trees of the entire dataset were constructed from the alignments, with each sequence colored to correspond to the sample time in (A). Trees were subjected to bootstrap analysis (1000 replicates). The branches having bootstrap support values >75% are highlighted in bold using the color of the sampling date. The outgroup in each case is pNL4-3; for ease of display the distance to the outgroup for some phylogenetic trees is reduced as indicated.

Figure 5.

Shifts in HIV populations with time. Plasma HIV RNA sequences were obtained from individual time points. The population subdivision test was performed for all pairwise combinations of samples for each patient dataset, and the probability of panmixia result is reported here as a function of the time between the sample pairs. Data for a series of 8 patients and 101 pairwise comparisons is presented. The median time to achieve a low probability of panmixia (10^{-9}) was 1017 days.

Figure 6.

Estimates of HIV replicating effective population size (N_e) *in vivo* using two methods. A. N_e was calculated for the virus population in each of the 10 patients shown as described in Materials and Methods using a coalescent-based method (diamonds). In addition, N_e was determined by measuring the change in allele frequencies for each polymorphic allele in *pro-pol*, and presented as box and whisker plots, with the box

JVI Accepts published online ahead of print

extending one quartile from the mean value and the ends of the whiskers indicating the extreme values. B. The population size estimated from changes in allele frequency at each individual site for all patients as a function of position in the *pro-pol* amplicon. Population sizes determined from allele frequency changes at synonymous (red) and nonsynonymous sites (blue) are indicated; box and whisker plots summarizing the average population size estimates for all patients are presented adjacent to the distribution.

JVI Accepts published online ahead of print

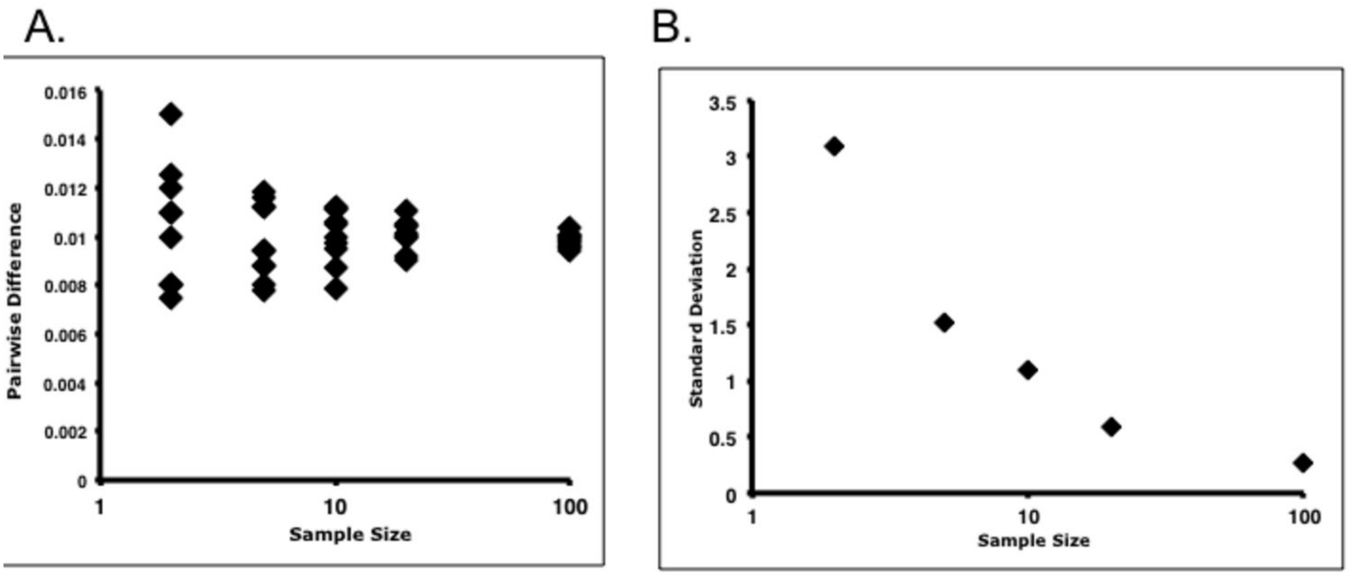


Figure 1

JVI Accepts published online ahead of print

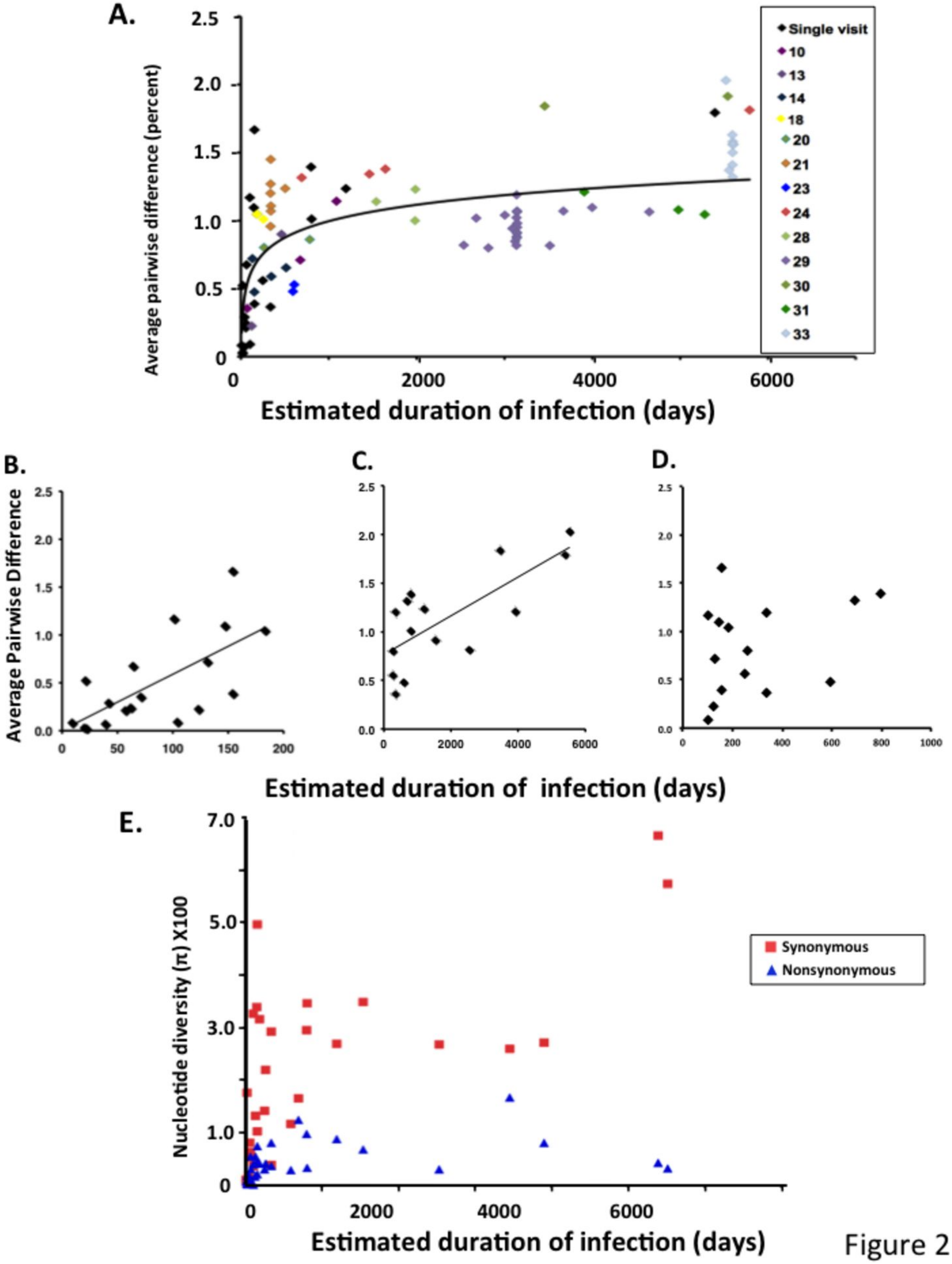


Figure 2

JVI Accepts published online ahead of print

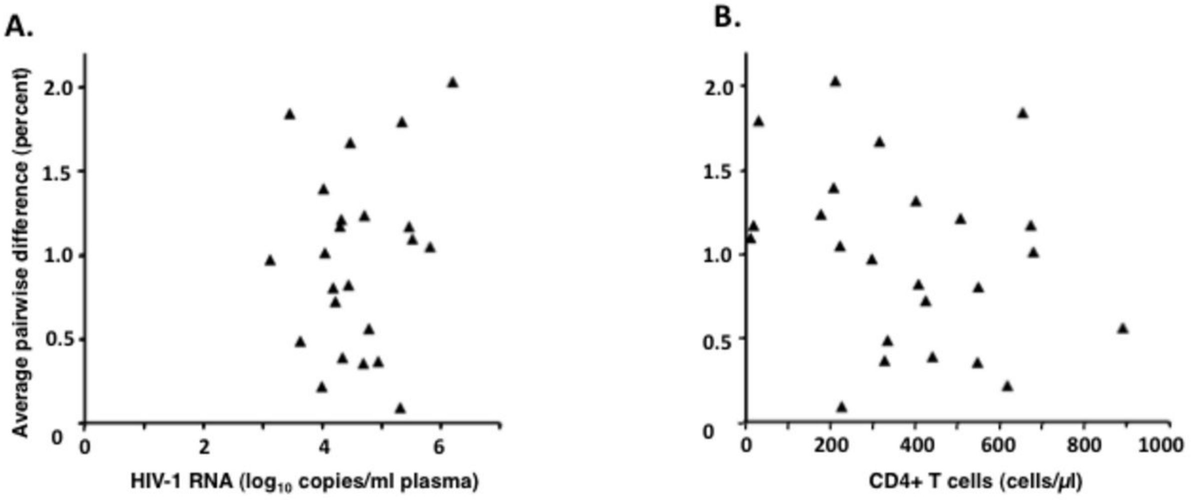


Figure 3

JVI Accepts published online ahead of print

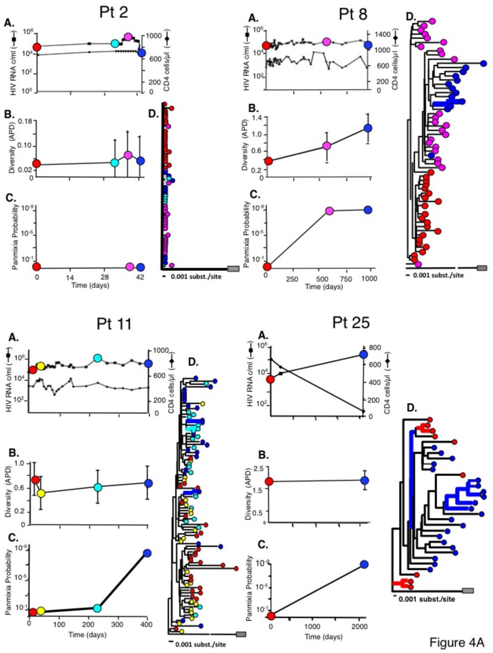


Figure 4A

JVI Accepts published online ahead of print

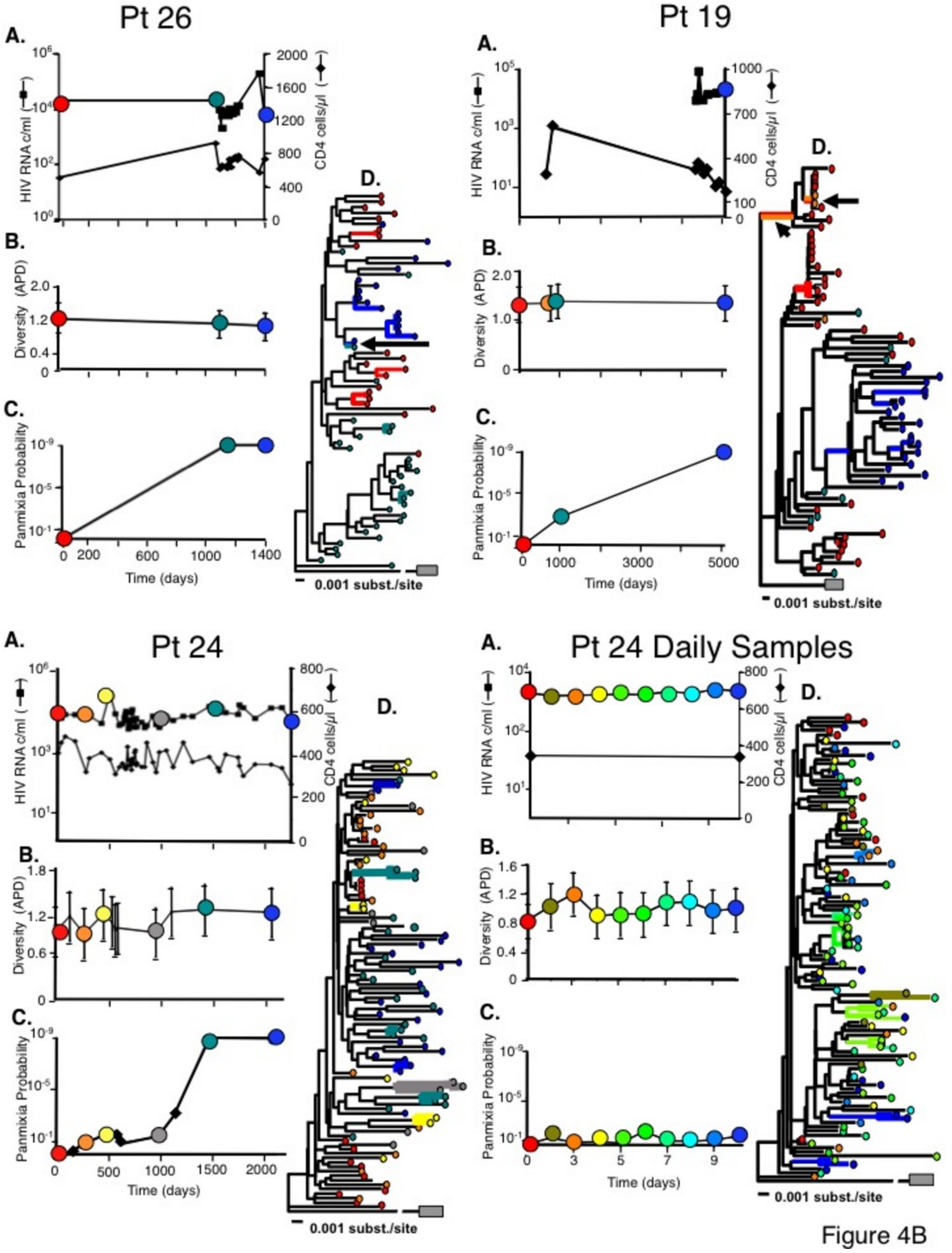


Figure 4B

JVI Accepts published online ahead of print

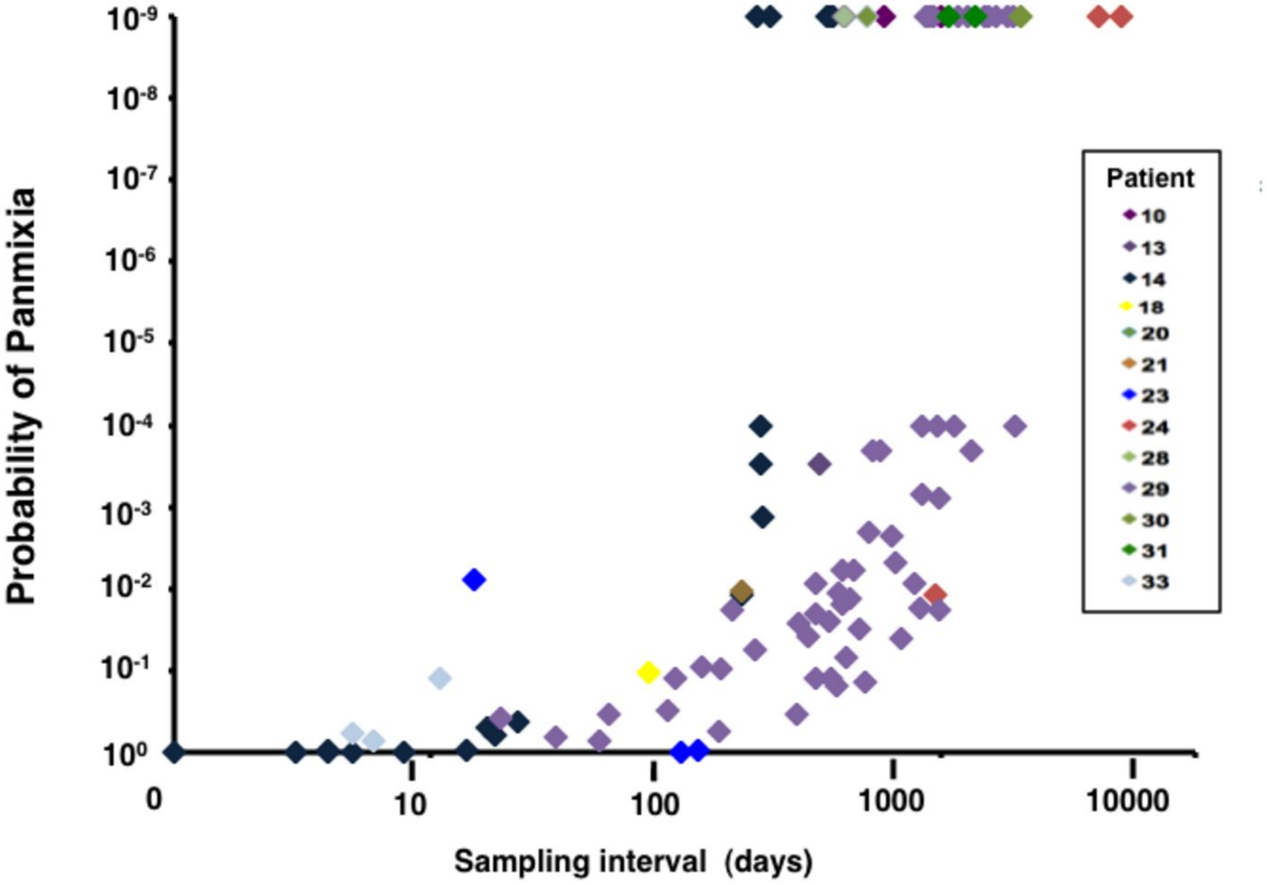


Figure 5

JVI Accepts published online ahead of print

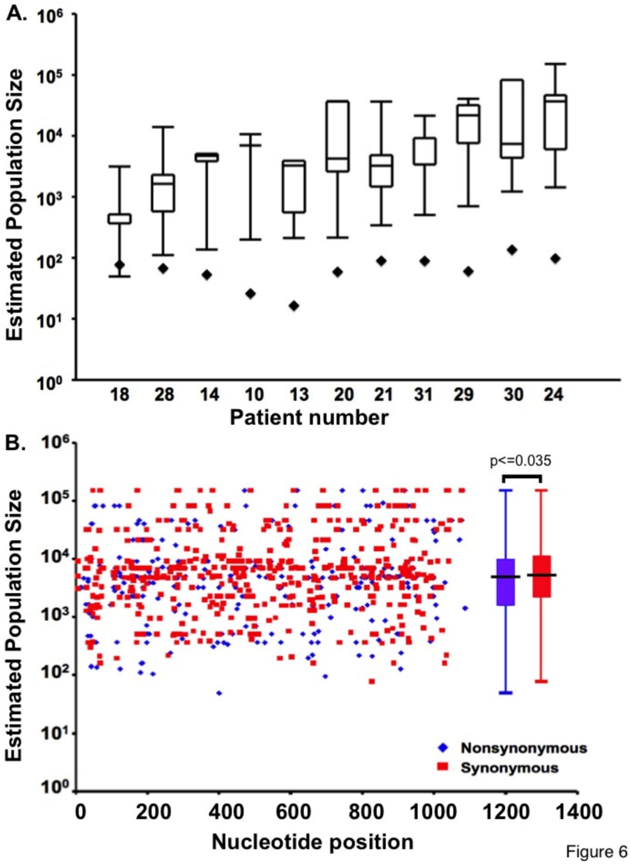
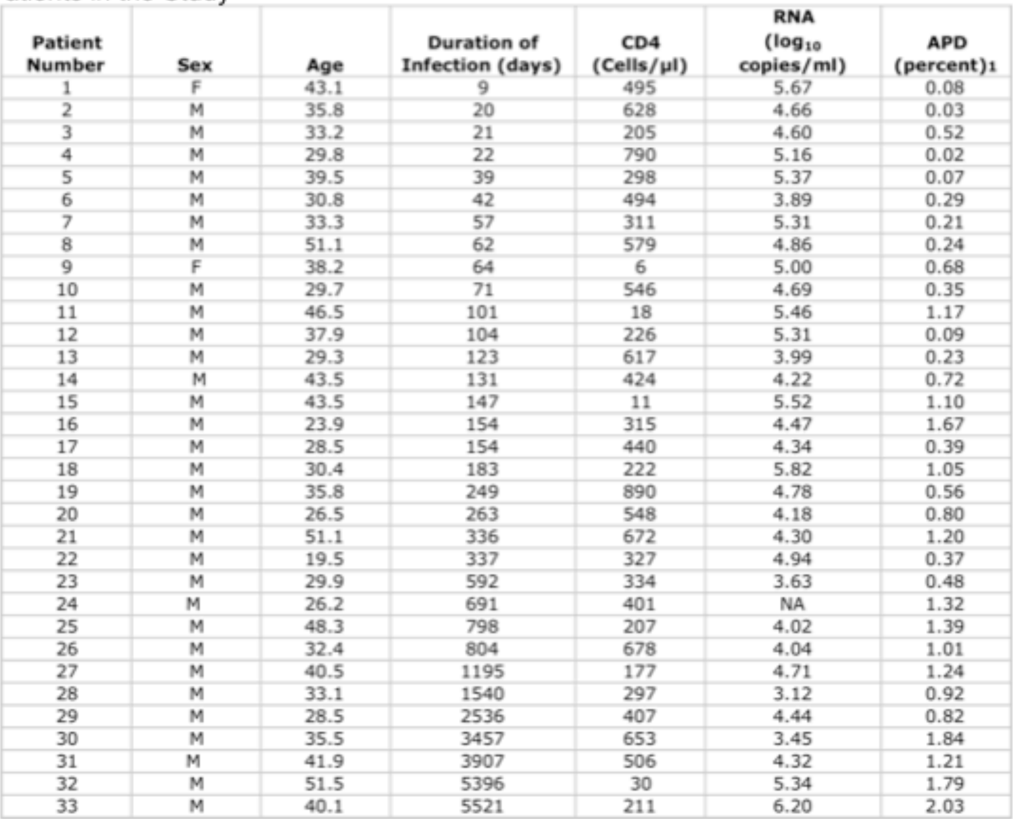


Figure 6

JVI Accepts published online ahead of print

Table 1. Patients in the Study



Patient Number	Sex	Age	Duration of Infection (days)	CD4 (Cells/µl)	RNA ($\log_{10}$ copies/ml)	APD (percent)[1]
1	F	43.1	9	495	5.67	0.08
2	M	35.8	20	628	4.66	0.03
3	M	33.2	21	205	4.60	0.52
4	M	29.8	22	790	5.16	0.02
5	M	39.5	39	298	5.37	0.07
6	M	30.8	42	494	3.89	0.29
7	M	33.3	57	311	5.31	0.21
8	M	51.1	62	579	4.86	0.24
9	F	38.2	64	6	5.00	0.68
10	M	29.7	71	546	4.69	0.35
11	M	46.5	101	18	5.46	1.17
12	M	37.9	104	226	5.31	0.09
13	M	29.3	123	617	3.99	0.23
14	M	43.5	131	424	4.22	0.72
15	M	43.5	147	11	5.52	1.10
16	M	23.9	154	315	4.47	1.67
17	M	28.5	154	440	4.34	0.39
18	M	30.4	183	222	5.82	1.05
19	M	35.8	249	890	4.78	0.56
20	M	26.5	263	548	4.18	0.80
21	M	51.1	336	672	4.30	1.20
22	M	19.5	337	327	4.94	0.37
23	M	29.9	592	334	3.63	0.48
24	M	26.2	691	401	NA	1.32
25	M	48.3	798	207	4.02	1.39
26	M	32.4	804	678	4.04	1.01
27	M	40.5	1195	177	4.71	1.24
28	M	33.1	1540	297	3.12	0.92
29	M	28.5	2536	407	4.44	0.82
30	M	35.5	3457	653	3.45	1.84
31	M	41.9	3907	506	4.32	1.21
32	M	51.5	5396	30	5.34	1.79
33	M	40.1	5521	211	6.20	2.03

[1] Average Pairwise Distance

JVI Accepts published online ahead of print

Table 2.
Polymorphism Analysis

Patient Number	HIV-1 RNA (copies/ml)	Duration of interval between sampling (days)	Polymorphisms with change in allele frequency (percent)*	Number of polymorphisms that arose or underwent fixation**
8	4.7	1017	3.6	3
10	4.0	339	14	4
11	4.2	383	5.9	0
13	5.8	72	0	0
15	4.2	520	10	3
16	4.3	168	8	1
19	NA	5099	43.3	21
23	3.1	422	13	0
24	4.4	2112	9	0
25	3.5	2085	5.1	3
26	4.3	1373	18	0
Median	4.3	520	9.0	1.0

* polymorphisms were identified and allele frequencies determined. Polymorphisms with change a significant change in allele frequency (Fisher's exact test $p<0.05$) were determined.

**determined as described in Methods.